

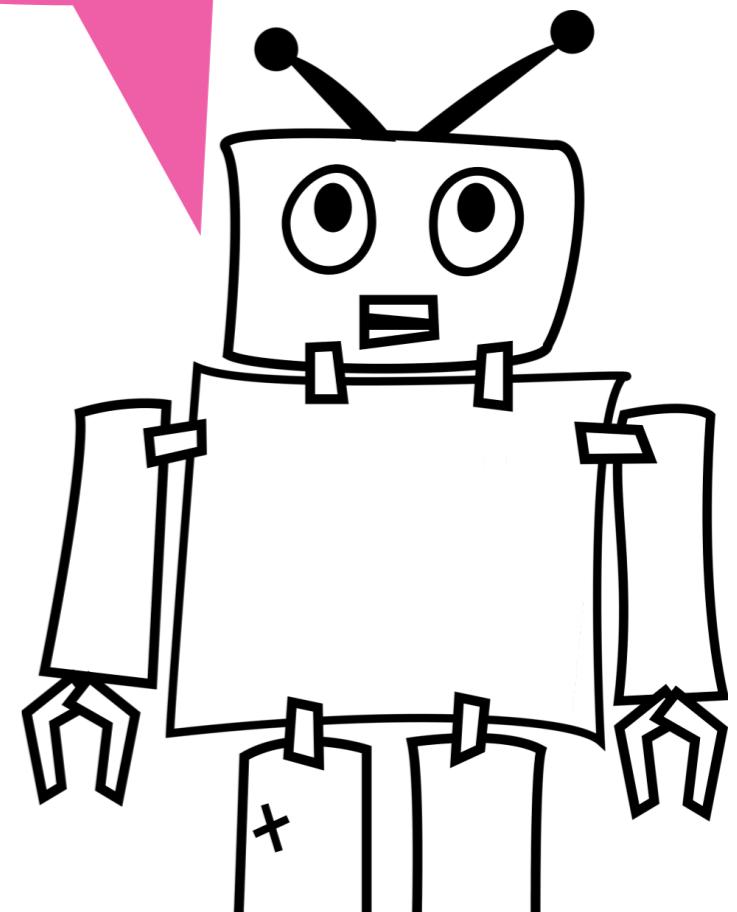
Question Answering and Reading Comprehension

Kevin Duh

Fall 2019, Intro to HLT, Johns Hopkins University

What is Question Answering?

It's a field concerned with building systems that answer questions posed in natural language



Question Answering (QA) vs. Information Retrieval (IR)

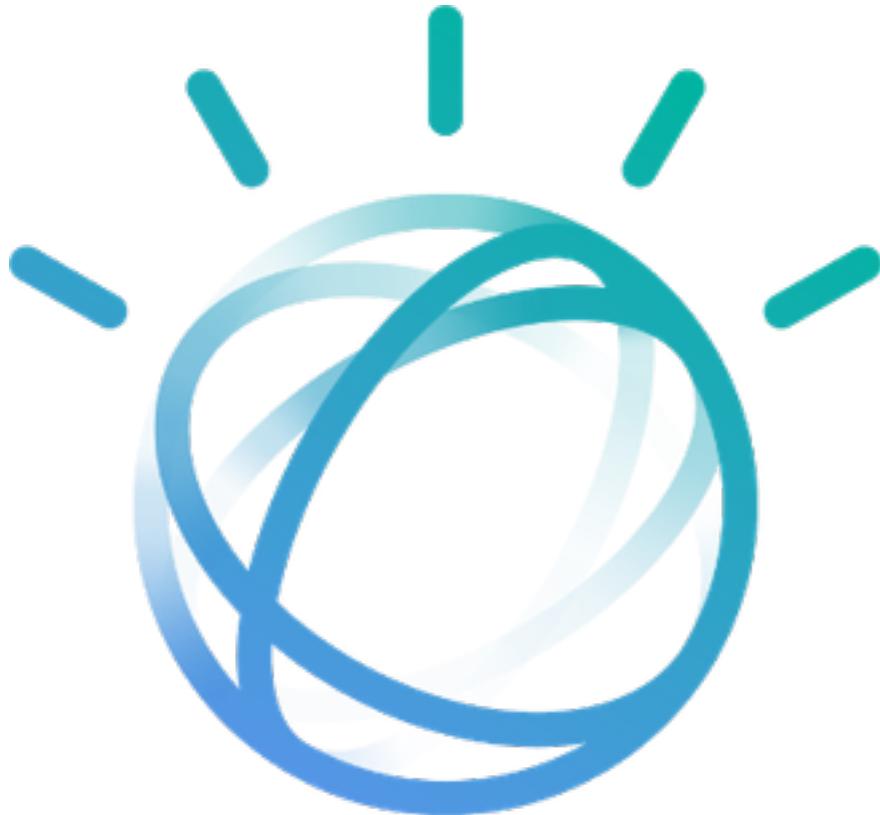
- QA and IR are related, but satisfy different info needs
- In QA, **questions are in natural language sentences**; in IR, queries tend to be short keyword phrases
- In QA, the **answers are often short and to-the-point**; in IR, the system returns lists of documents.
- In QA, the **answer might be synthesized from multiple sources**; In IR, a document is the atomic unit.

QA systems integrate many HLT technologies

- Building a QA system is like doing a **triathlon**. You need to be good at many things, e.g.
 - Parsing, Information Extraction, Semantic Role Labeling, Knowledge Bases, Supervised/Semi-supervised learning, Distributed Processing, Information Retrieval...



IBM Watson wins on *Jeopardy!* Quiz Show (2011)



https://commons.wikimedia.org/wiki/File:IBM_Watson_w_Jeopardy.jpg

- See it in action:
 - <https://www.youtube.com/watch?v=P18EdAKuC1U>
 - https://www.youtube.com/watch?v=WFR3lOm_xhE

Outline

- Question Answering (QA)
 - Problem Formulation
 - System architecture (an example)
- Machine Reading Comprehension (MRC)
 - Problem Formulation
 - System architecture (an example)
- Future Directions

Question Types

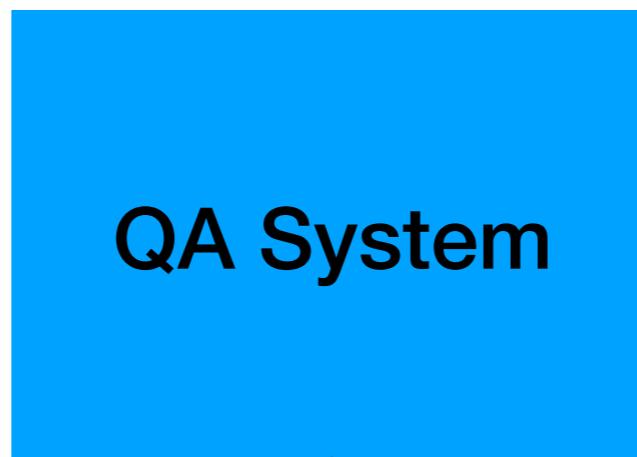
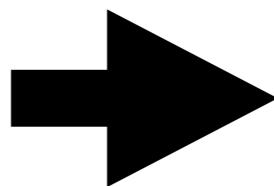
- Factoid Question: Who was the first American in space?
Alan Shepard
- List Question: Name 20 countries that produce coffee
Brazil, Vietnam, Colombia, Indonesia, Ethiopia, Honduras, India, Uganda, ...
- Definition Question: Who is Aaron Copland?
He is an American composer, composition teacher, writer, and conductor. His best-known works in 1930s and 1940s include Appalachian Spring, Rodeo, ...
- Relationship Question: Are Israel's military ties to China increasing?
Yes (arms deal ~1993). Now, it's more complex to answer this. There's strengthening of investments/trade, and delicate relation w.r.t. the U.S.
- Opinion Question: Why do people like Trader Joe's?
Friendly employees, maybe?

QA Challenges

- Flexibility and ambiguity of human language makes it challenge to match question to answer-bearing text
- Answer may differ depending on time
 - Q: *Which car manufacturer is owned by VW since 1998?*
 - Candidate text in 1993: Volkswagen today announced the acquisition of Bently
- Answer may need synthesizing multiple sources or reasoning
 - Q: *In which country is Sony headquartered?*
 - We have evidence it's in Tokyo. And Tokyo is a city in Japan.

Problem Formulation

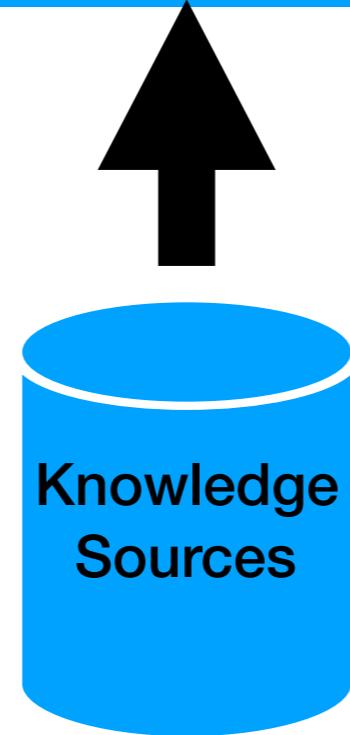
Question



Answer

Usually, we'll restrict the question type for each task

We'll assume factoid questions for the rest of these slides. (It's been most investigated)



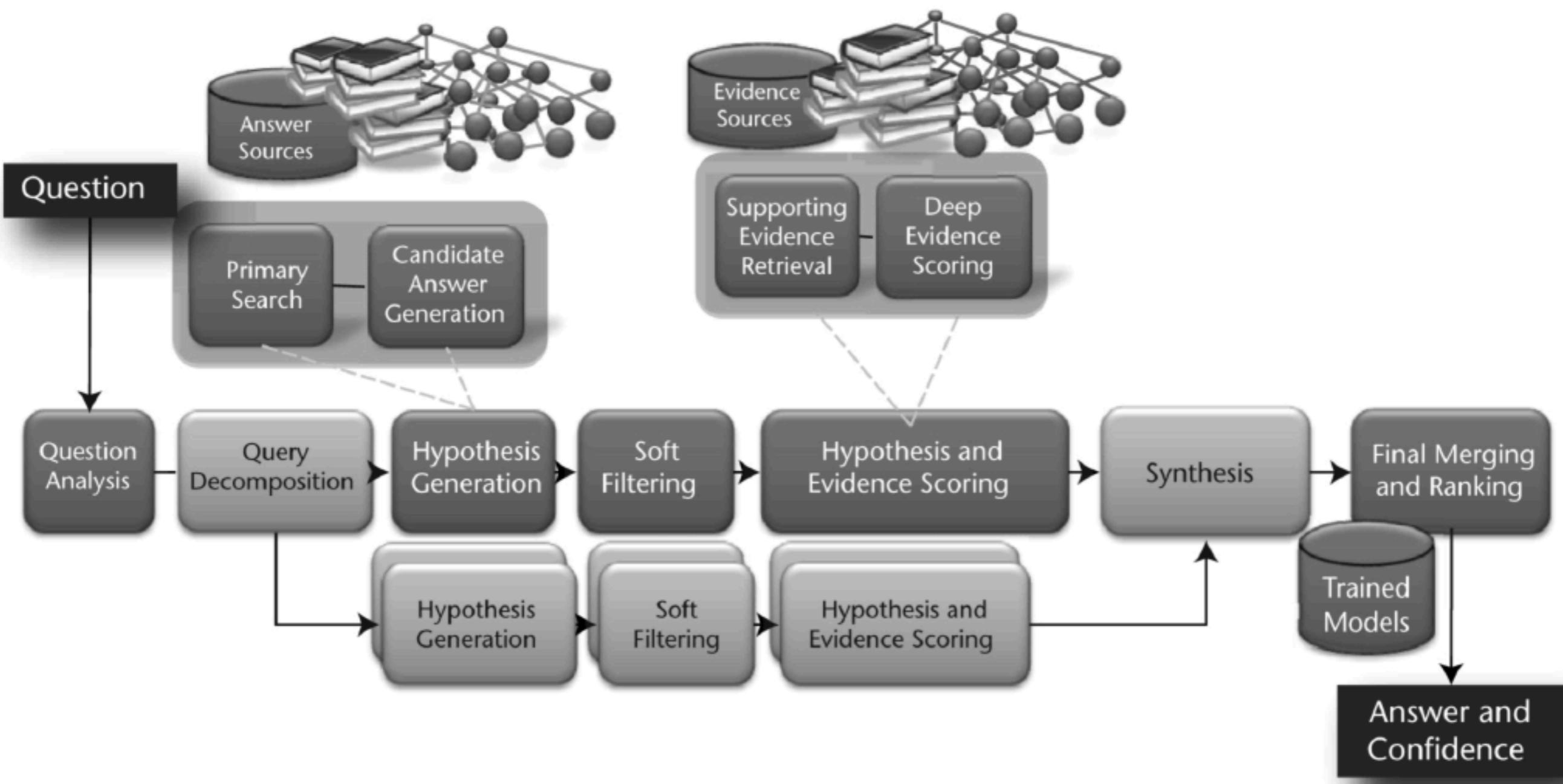
Evaluation metrics include:

- Accuracy
- Rank-based metrics (*MRR*)
- Precision/Recall/F-score
- Confidence-weighted metric

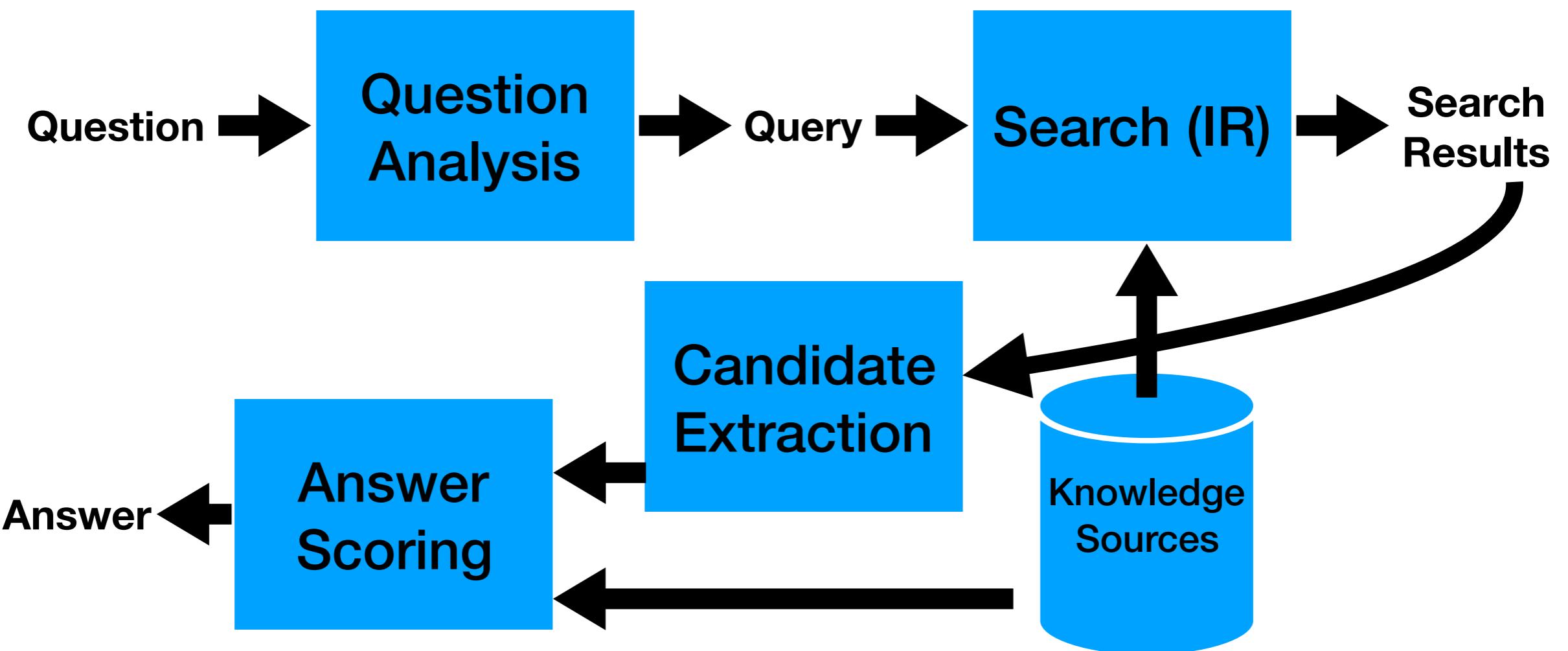
Outline

- Question Answering (QA)
 - Problem Formulation
 - System architecture (an example)
- Machine Reading Comprehension (MRC)
 - Problem Formulation
 - System architecture (an example)
- Future Directions

IBM Watson Architecture for Jeopardy!

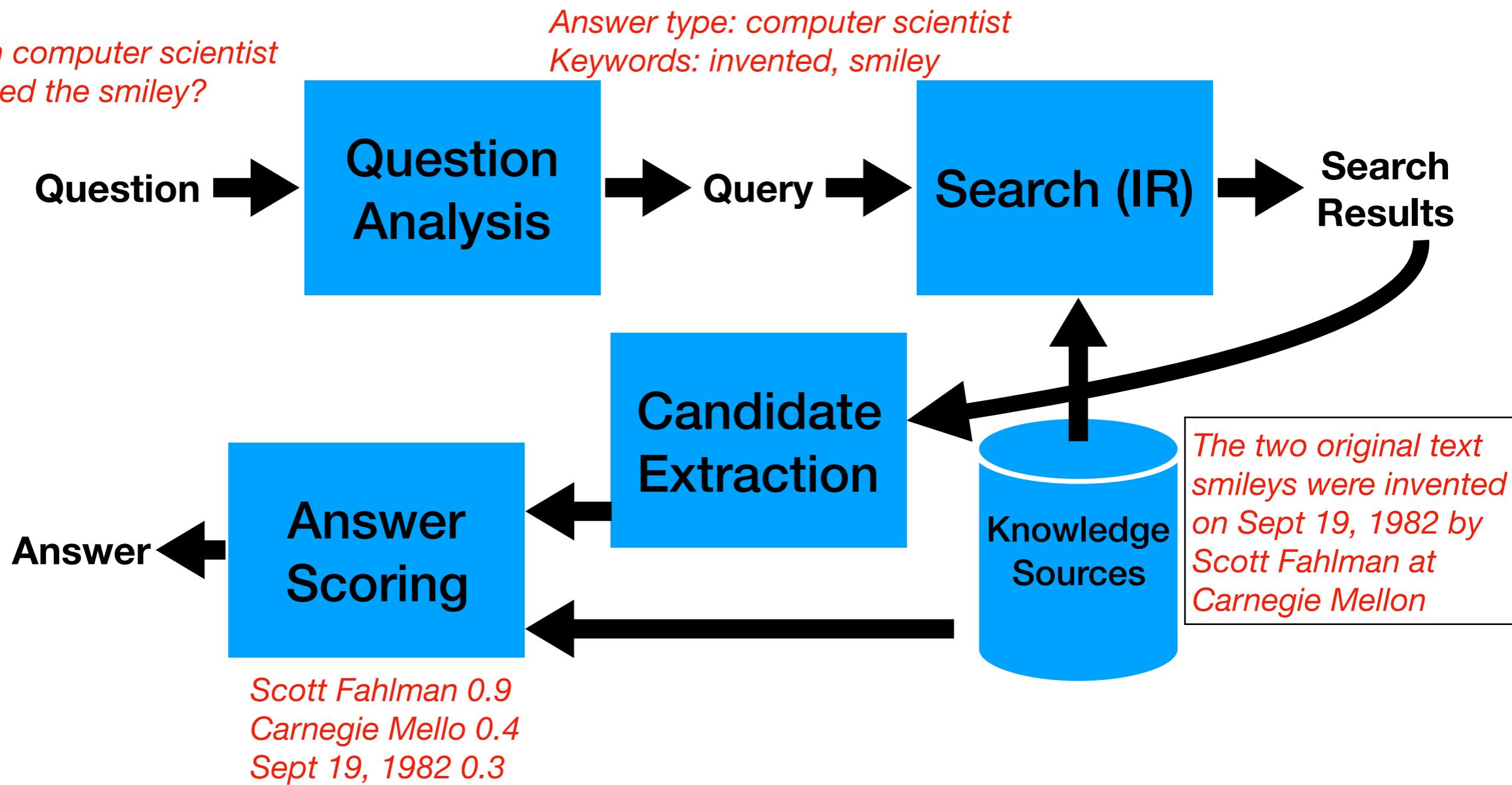


We'll discuss a simpler but similar architecture



We'll discuss a simpler but similar architecture

Which computer scientist invented the smiley?



Question Analysis

- It's important to get the answer type
 - Q: Who invented the light bulb? Type: PERSON
 - Q: How many people live in Bangkok? Type: NUMBER
- Answer type labels are usually arranged in an ontology to address answers of different granularities
- Answer type classifier could be regex, or machine learned system based on answer type and question pairs

Search

- **Keyword query** (e.g. using informative words from question) is often used.
 - Exploits IR advances, e.g. query expansion
- **Structured query** with more linguistic processing helps:
 - named entity recognition, relation extraction, anaphora
- Return documents, then split into passages. Or directly work with indexed passages.

Candidate Extraction

- A mixture of approaches, based on answer type result
- Exhaustive list of instances in a type:
 - e.g. the names all U.S. presidents, regex for numbers
 - high recall, but assume valid type
- Syntactic/Semantic matching of question & candidate
 - Q: *Who killed Lee Harvey Oswald?* Answer type: PERSON
 - Text: Kennedy was killed by Oswald.
 - What should be the answer candidates? Kennedy, Oswald, or neither?
 - Semantic roles will improve precision, but computationally expensive

Answer Scoring

- Knowledge source might be redundant, containing multiple instances of the same candidate answer
 - Multiple evidence increases confidence of answer
 - Candidates may need to be normalized before evidence combination. e.g. “Rome, Italy” vs “Rome”.
- We may also have candidate answers from databases rather than text sources
- Often uses machine learning to integrate many features

Outline

- Question Answering (QA)
 - Problem Formulation
 - System architecture (an example)
- Machine Reading Comprehension (MRC)
 - Problem Formulation
 - System architecture (an example)
- Future Directions

Machine Reading Comprehension (MRC) Task

In meteorology, precipitation is any product of the condensation of atmospheric water vapor that falls under gravity. The main forms of precipitation include drizzle, rain, sleet, snow, graupel and hail. Precipitation forms as smaller droplets coalesce via collision with other rain drops or ice crystals within a cloud.

Question:

What causes precipitation to fall?

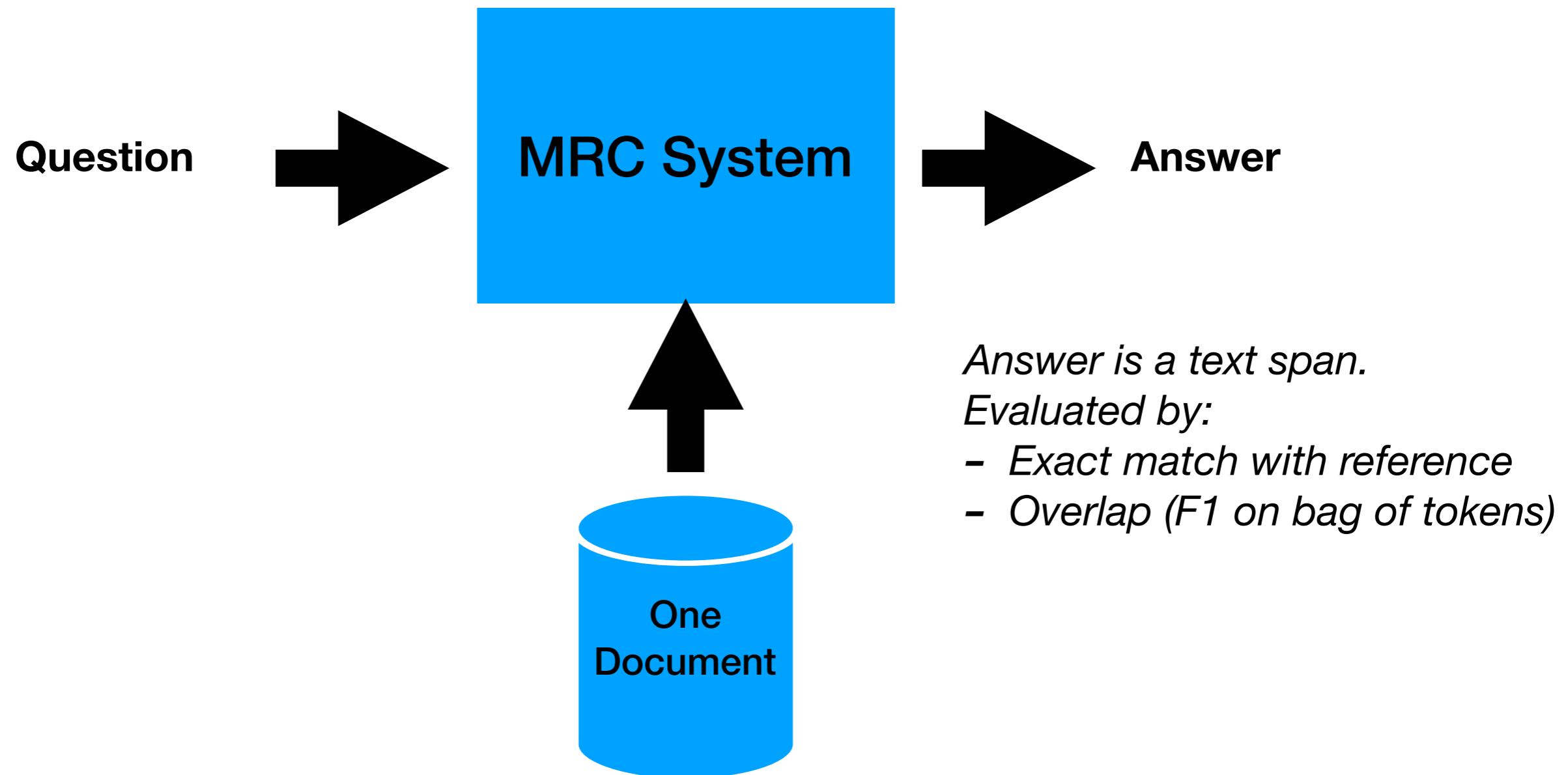
Answer: **gravity**

Question:

What is another main form of precipitation besides drizzle, rain, snow, sleet and hail?

Answer:

Problem Formulation (as in SQuAD v1.0)



MRC vs QA

- MRC tasks are designed to test the capabilities of reading and reasoning. QA focuses more on end-user.
- MRC is usually restricted to one document where the answer is present, to be read in depth; QA exploits multiple knowledge sources.

Question types in SQuAD

Reasoning	Description	Example
Lexical variation (synonymy)	Major correspondences between the question and the answer sentence are synonyms.	Q: What is the Rankine cycle sometimes called ? Sentence: The Rankine cycle is sometimes referred to as a <u>practical Carnot cycle</u> .
Lexical variation (world knowledge)	Major correspondences between the question and the answer sentence require world knowledge to resolve.	Q: Which governing bodies have veto power? Sen.: <u>The European Parliament and the Council of the European Union</u> have powers of amendment and veto during the legislative process.
Syntactic variation	After the question is paraphrased into declarative form, its syntactic dependency structure does not match that of the answer sentence even after local modifications.	Q: What Shakespeare scholar is currently on the faculty ? Sen.: Current faculty include the anthropologist Marshall Sahlins, ..., Shakespeare scholar David Bevington.
Multiple sentence reasoning	There is anaphora, or higher-level fusion of multiple sentences is required.	Q: What collection does the V&A Theatre & Performance galleries hold? Sen.: <u>The V&A Theatre & Performance galleries</u> opened in March 2009. ... They hold the UK's biggest national collection of material about live performance.
Ambiguous	We don't agree with the crowd-workers' answer, or the question does not have a unique answer.	Q: What is the main goal of criminal punishment? Sen.: Achieving crime control via <u>incapacitation and deterrence</u> is a major goal of criminal punishment.

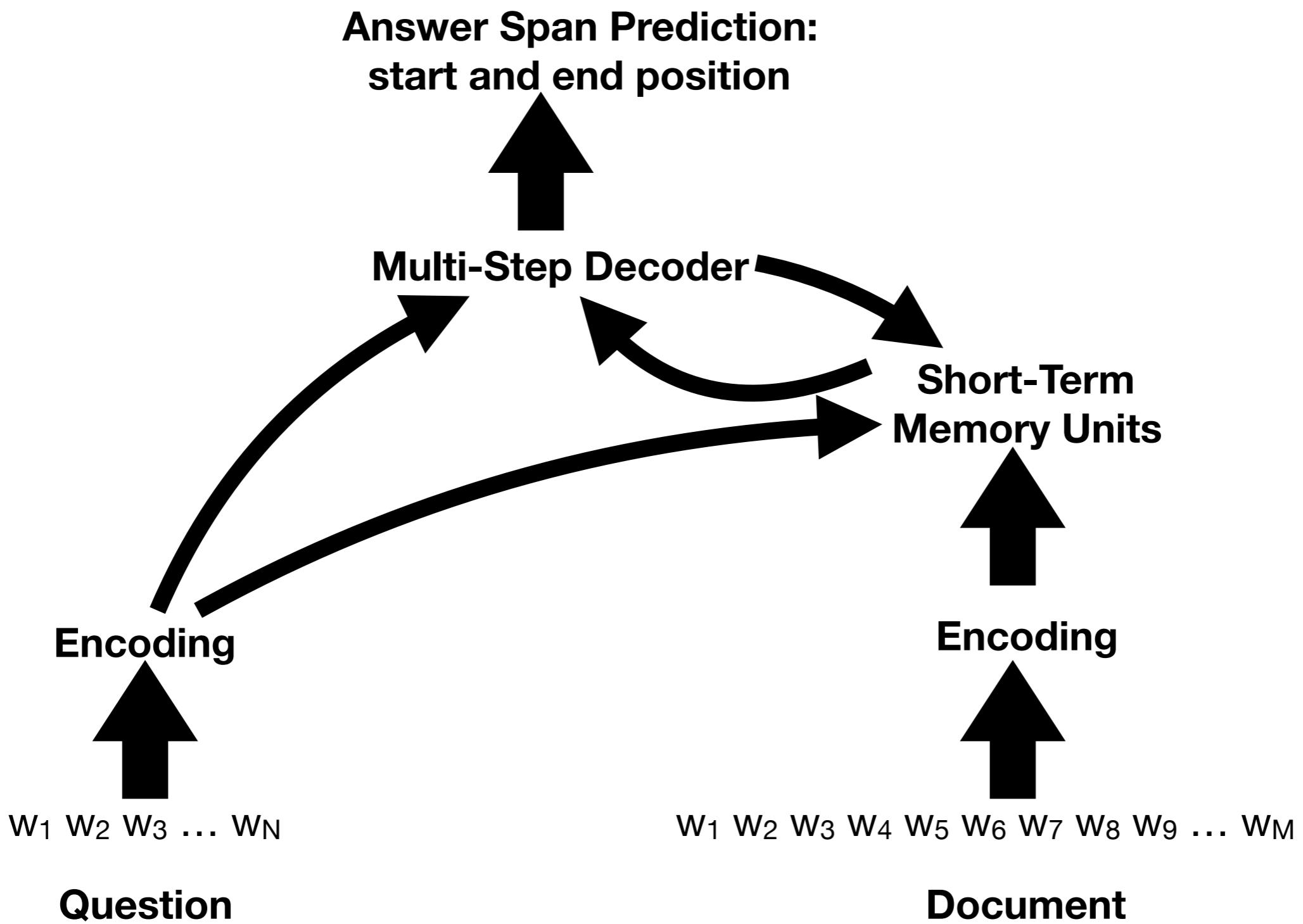
Outline

- Question Answering (QA)
 - Problem Formulation
 - System architecture (an example)
- Machine Reading Comprehension (MRC)
 - Problem Formulation
 - System architecture (an example)
- Future Directions

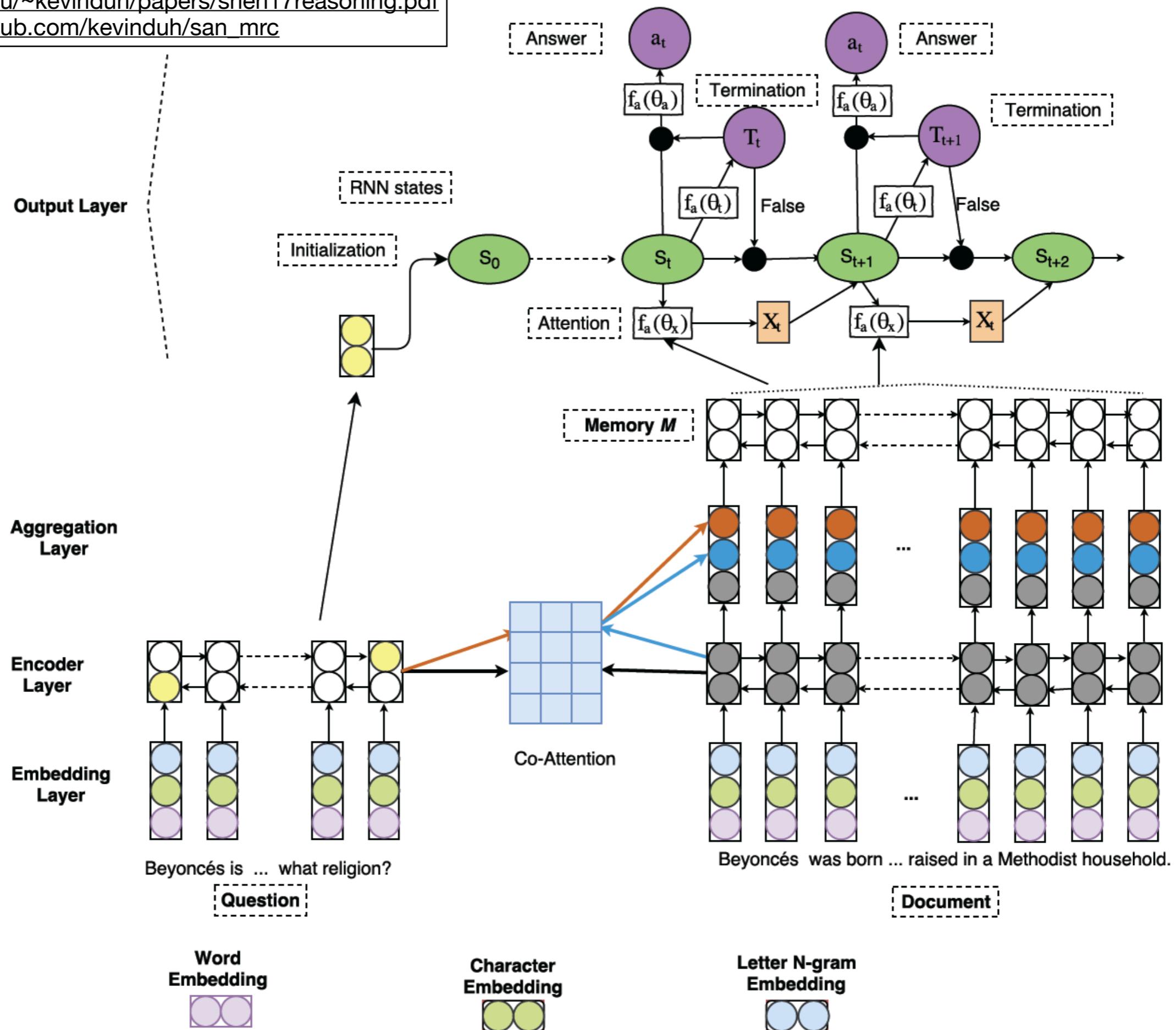
Multi-Step Reasoning

- Question: What collection does the V&A Theater & Performance galleries hold?
- Document: The V&A Theater & Performance galleries opened in March 2009. ... They hold the UK's biggest national collection of material about live performance.
- Answer in multi-step:
 - Perform coference resolution to link “They” and “V&A”
 - Extract direct object from “They hold ___”

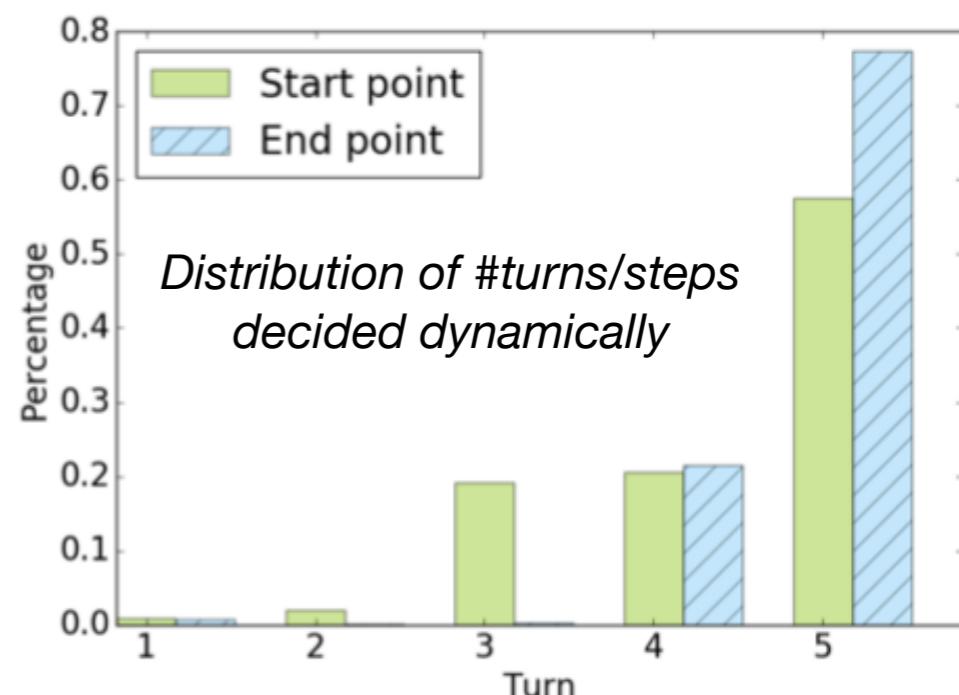
A Neural Model Architecture



From: Liu et. al. (2017) An Empirical Analysis of Multiple-Turn Reasoning Strategies in Reading Comprehension Tasks.
<http://www.cs.jhu.edu/~kevinduh/papers/shen17reasoning.pdf>
See also: https://github.com/kevinduh/san_mrc



Example Run



P: Forces act in a particular direction and have sizes dependent upon how strong the push or pull is. Because of these characteristics, forces are classified as “vector quantities”... For example, when determining what happens when two forces act on the same object, it is necessary to **know both the magnitude and the direction of both forces to calculate the result**. If both of these pieces of information are not known for each force, the situation is ambiguous... Associating forces with vectors avoids such problems.

Q: How do you avoid problems when determining forces involved on an object from two or more sources?

(Turn 1): when determining what happens when two forces act on the same object

(Turn 2): two forces act on the same object

(Turn 3): it is necessary to know both the magnitude and the direction of both forces

Outline

- Question Answering (QA)
 - Problem Formulation
 - System architecture (an example)
- Machine Reading Comprehension (MRC)
 - Problem Formulation
 - System architecture (an example)
- Future Directions

Many active areas of research!

(Part of larger push on Artificial Intelligence)

- New benchmarks and data curation methods
 - Text-based: SQuAD 2.0, MS MARCO, ...
 - Multi-modal: Visual QA
 - Incorporating Commonsense Reasoning
- Grand Challenges: Todai Robot, AI2 Aristo, etc.
- New opportunities in applications
 - Customer service chatbot, Siri digital assistant, Watson health

Visual QA

<https://visualqa.org>

Question : Why are the men jumping?

Original Image | **to catch frisbee**



Complementary Image | **trick**



Commonsense Reasoning

- Commonsense about the physical world
 - e.g. [Winograd Schema Challenge (Levesque 2011)]
 - Q: The trophy would not fit in the brown suitcase because it was too big. What was too big?
 - A. The trophy
 - B. The suitcase

See: Storks, Gao, Chai (2019). Commonsense Reasoning for Natural Language Understanding: A Survey of Benchmarks, Resources, and Approaches <https://arxiv.org/pdf/1904.01172.pdf>

Commonsense Reasoning

- Commonsense about the social world
 - e.g. [ROCStories (Mostafazadeh et al., 2016)]
 - Tom and Sheryl have been together for two years. One day, they went to a carnival together. He won her several stuffed bears, and bought her funnel cakes. When they reached the Ferris wheel, he got down on one knee. [Finish the story]
 - A. Tom asked Sheryl to marry him.
 - B. He wiped mud off of his boot.

See: Storks, Gao, Chai (2019). Commonsense Reasoning for Natural Language Understanding: A Survey of Benchmarks, Resources, and Approaches <https://arxiv.org/pdf/1904.01172.pdf>

AI2 Project Aristo: solving elementary/middle school science exams

Which object is the best conductor of electricity?

- (A) metal fork
- (B) rubber boot
- (C) plastic spoon
- (D) wooden bat

Aristo's Answer: (A) metal fork

Correct Answer: A

Confidence: 90.72%

as computed from these reasoners:

Information Retrieval: 57.12% [MORE INFO](#)

Justification Sentence: For metals, the thermal conductivity is quite high, and those metals which are the best electrical conductors are also the best thermal conductors.

Table Reasoning: 38.90% [MORE INFO](#)

Knowledge Used: [metal | conductor | electricity]

Topic Matching: 52.72% [MORE INFO](#)

Topic: conductor, conductors

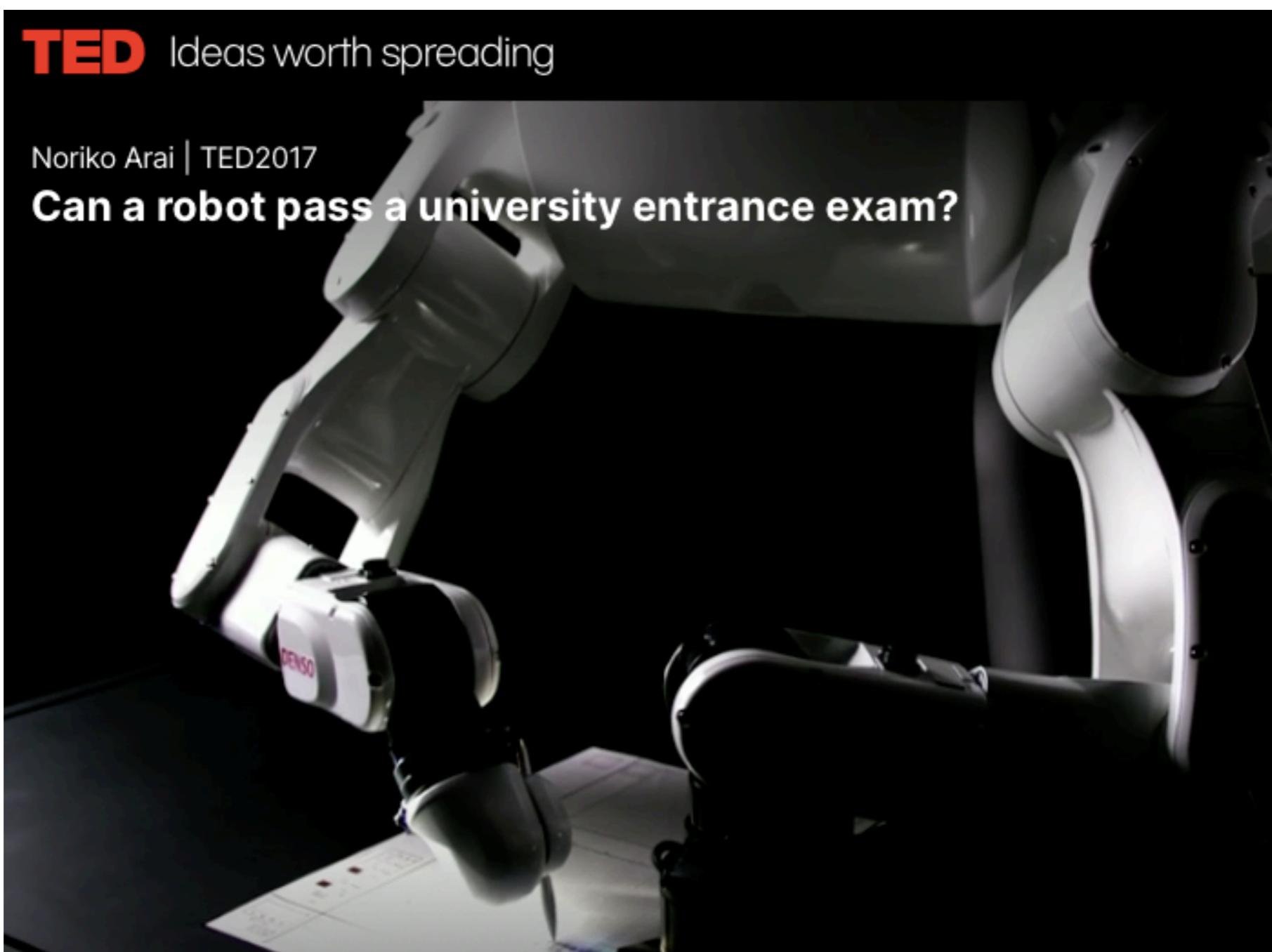
Tuple Reasoning: 46.23% [MORE INFO](#)

Knowledge Used: [all metal objects | are | conductors of electricity] [metal | type Of | conductor | electricity] [conductor | object | electricity | metal] [metal | type Of | conductor]

AristoRoBERTa: 90.30% [MORE INFO](#)

Noriko Arai | TED2017

Can a robot pass a university entrance exam?



https://www.ted.com/talks/noriko_arai_can_a_robot_pass_a_university_entrance_exam/

Todai Robot Project (2011-2016)

- passing an exam requires multiple intelligences
- what's easy for human may not be easy for computers, and vice versa
- multiple choice, written essay questions
- topics: social studies, math, physics, English, Japanese
- https://www.nii.ac.jp/userdata/results/pr_data/NII_Today/60_en/all.pdf

Summary

- Question Answering (QA)
 - Problem Formulation
 - System architecture (an example)
- Machine Reading Comprehension (MRC)
 - Problem Formulation
 - System architecture (an example)
- Future Directions

Questions?